АЛГОРИТМ ВЫСОКОТОЧНОЙ РАЗМЕТКИ НА ПИТЧИ ЭЛЕМЕНТОВ КОМПИЛЯЦИИ ДЛЯ СИНТЕЗА РЕЧИ ПО ТЕКСТУ[1]

A HIGH PRECISION PITCH MARKER OF COMPILATION UNITS FOR TEXT-TO-SPEECH SYNTHESIS

Лобанов Б.М (lobanov@newman.bas-net.by), Давыдов А.Г. (andrew@ssrlab.com)

Объединенный институт проблем информатики НАН Беларуси, Минск, Беларусь

Описывается адаптивный алгоритм высокоточной разметки на питчи баз данных речевых элементов для систем компиляционного синтеза речи по тексту. Приведены программная реализация и результаты тестирования на мужских и женских голосах.

Введение

Высокоточная разметка речевого сигнала (РС) по его осциллограмме является непростой задачей даже для опытного фонетиста из-за нестационарности и чрезвычайной изменчивости речи. Между тем, точная и безошибочная разметка на периоды основного тона (питчи) элементов компиляции является непременным условием синтеза неискаженной речи при её просодической модификации. При этом желательно, чтобы маркеры питчей точно указывали моменты смыкания голосовых связок во время фонации. Это необходимо для правильного функционирования SL-алгоритма [1] изменения частоты основного тона (ЧОТ) при интонировании речи. SL-алгоритм, в отличие от широко используемого алгоритма PSOLA [2], позволяет осуществлять “щадящую” модификацию ЧОТ путём “плавной сшивки” (“Soft Lacing”) соседних периодов естественного сигнала на интервалах открытой голосовой щели, сохраняя речевой сигнал неизменённым на остальных участках.

Существует множество алгоритмов определения ЧОТ речевого сигнала [3–5], однако ни один из них не удовлетворил нас по совокупному критерию точности и надёжности при одновременном выполнении требования алгоритмической простоты. Данная работа посвящена постановке и решению такой задачи.

1. Адаптивный алгоритм оценки мгновенных значений ЧОТ

Один из наиболее эффективных, по критерию вычислительной сложности, методов определения ЧОТ и меры периодичности речевого сигнала [6] базируется на вычислении сдвиговой функции (СФ):

$$\gamma_n(k)=\frac{1}{M}\sum_{m=0}^{M-1}\left|x(n+m)-x(n+m-k)\right|,$$

где $x(n)$ – анализируемый речевой сигнал (РС); M – интервал наблюдения СФ в отсчётах.

Для вокализованных участков РС функция $\gamma_n(k)$ будет иметь глубокие провалы при величине задержки $k=\pm T_0,\pm 2T_0,\cdots$, где T_0 – период основного тона (ОТ). Интервал наблюдения M целесообразно выбирать таким образом, чтобы он включал хотя бы один период ОТ. Примеры СФ, вычисленные для вокализованного и невокализованного участков РС, приведены на рис. 1.

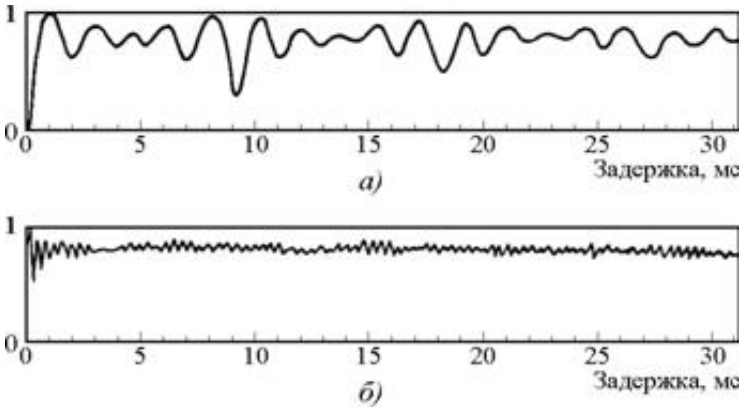


Рис.1. Нормированная СФ: (а) – для вокализованного и (б) – для невокализованного участков речи

Определение текущего значения ЧОТ осуществляется путём нахождения минимумов СФ, соответствующих

периоду ОТ. При этом для РС возможны ошибки 2-х типов: нахождение ложного минимума, соответствующего удвоенному периоду ОТ (2-х кратное уменьшение реального значения ЧОТ) или минимума, соответствующего периоду колебаний 1-й форманты (вплоть до 2-х кратного увеличения реального значения ЧОТ).

Для устранения ложных скачков траектории ЧОТ в [6] рассматриваются такие нелинейные методы коррекции как центральное ограничение автокорреляционной функции и медианное сглаживание в комбинации с линейным сглаживанием ЧОТ. Однако из-за того, что при проведении коррекции не выполняется разделение речевого сигнала на вокализованные и невокализованные участки вероятность возникновения указанного рода ошибок остаётся ещё достаточно высокой. Для преодоления указанных трудностей предлагается использовать следующий адаптивный алгоритм определения ЧОТ.

1. Предполагая, что частота основного тона расположена в диапазоне от $F_{0\min} = 60$ Гц до $F_{0\max} = 500$ Гц, исходный РС пропускается через полосовой фильтр с полосой пропускания равной принятому диапазону частоты основного тона.

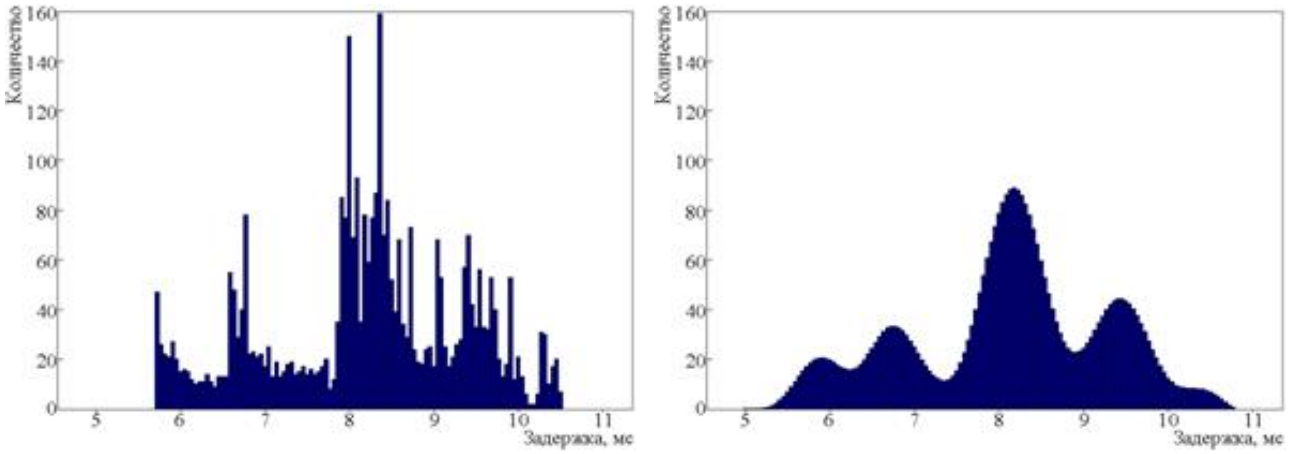
2. На всём отфильтрованном РС определяется текущая мера его вокализованности (мера тона), оцениваемая по величине минимума нормированной СФ при значении задержки от $1/F_{0\max}$ до $1/F_{0\min}$:

$$\mu_n = 1 - \min\left(\tilde{y}_n(k)\right), \text{ для } k = \left[\frac{1}{F_{0\max}}, \frac{1}{F_{0\min}}\right].$$

3. Оцениваются текущие значения периода ОТ как значение задержки, при котором было определено минимальное значение СФ.

4. Для вокализованных участков РС (регионов), где $\mu_n \geq 0{,}6$ рассчитывается статистическое распределение значений периода ОТ и после его сглаживания определяется модальное значения (рис. 2).

5. Для устранения ошибок определения значений периода ОТ используется процедура адаптивной коррекции полученных мгновенных значений периода ОТ, путем переоценки СФ с учетом модального значения сглаженного распределения периода ОТ на вокализованном регионе.



а) **б)**

Рис. 2. Пример а) исходного и б) сглаженного распределений периода ОТ

6. Для адаптивной коррекции СФ предлагается воспользоваться следующей формулой:

$$\tilde{y}_n(k) = \left(\tilde{y}_n(k) - 1\right) \cdot e^{\alpha(k-\tau)^2} + 1, \quad \alpha = \frac{4\ln(s)}{\tau^2},$$

где $\tilde{y}_n(k)$ – исходная СФ; $\tilde{y}_n(k)$ – скорректированная СФ; τ – мода, найденная по усредненному распределению периода основного тона вокализованного региона; s – величина от 0 до 1, определяющая крутизну корректирующей функции, как ее значение в точке $\tau/2$.

Пример результата коррекции с использованием функции $\tilde{y}_n(k) = 0{,}1 \cdot \sin(0{,}1 \cdot k) + 0.9$, $\tau = 250$ и $s = 0{,}7$ приведен на рис. 3.

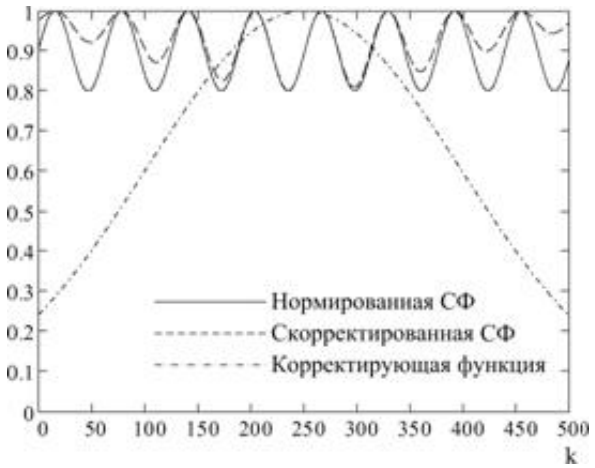


Рис. 3. Пример коррекции нормированной СФ

7. Заново оцениваются значения периодов ОТ и рассчитывается результирующая траектория частоты основного тона для всех вокализованных регионов с учетом проведенной коррекции.

2. Алгоритм расстановки питчей

На основе полученных оценок мгновенных значений ЧОТ расставляются маркеры питчей РС. Для того, чтобы маркеры питчей указывали моменты смыкания голосовых связок во время фонации, необходимо вначале определить «полярность» асимметрии РС. Асимметрия РС возникает в связи с тем, что первая полуволна затухающих формантных колебаний, возбуждаемых резким смыканием голосовых связок, как правило, имеет наибольшую амплитуду на периоде ОТ и может являться индикатором местоположения питча.

Для определения «полярности» речевого сигнала (в зависимости от положения микрофона при записи импульсы основного тона могут быть направлены в положительную либо в отрицательную область) для всех вокализованных регионов вычисляется разность положительных и отрицательных максимальных значений РС. В соответствии с определенной полярностью отыскивается положения локальных экстремумов на интервалах времени, соответствующих длительности периодов ОТ.

Питчи в каждом вокализованном регионе проставляются последовательно от положения экстремума к границам на основе изменения скорректированной траектории ОТ, с уточнением положения следующего локального экстремума на сигнале x в окрестности $\pm T_{0\max} \cdot T_{0i}$, где $T_{0\max}$ – максимально допустимое изменение периода ОТ от одного питча к другому (рабочее значение которого принято равным 0,15), T_{0i} – значение периода основного тона в i-ой позиции.

Для установки питчей в моменты смыкания голосовых связок необходимо для каждого найденного экстремума найти ближайшую слева позицию пересечения отфильтрованного сигнала $x(n)$ с нулем и в заключение уточнить данную позицию питча на исходном сигнале. Пример расстановки питчей на РС показан на рис. 4.

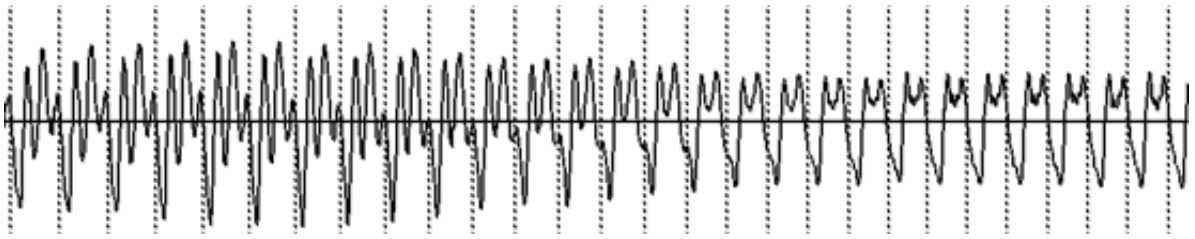


Рис. 4. Пример разметки вокализованного региона на периоды ОТ (питчи)

3. Программная реализация и тестирование системы

Параметры диалога настроек программного модуля разметки речевого сигнала на питчи (рис. 5) сгруппированы в блоки в соответствии с последовательностью выполнения операций:

1. выбор низкочастотного фильтра для выделения частотного диапазона ОТ;
2. задание настроек вычисления СФ функции (значения данного блока задаются в отсчетах сигнала с частотой дискретизации 22050 Гц.) и коэффициента s корректирующей функции;
3. задание параметров для нахождения вокализованных регионов;
4. выбора метода определения полярности асимметрии речевого сигнала;
5. выбора позиции расстановки питчей и сигнала, на котором будут расставляться питчи;
6. выбор предельного значения изменения частоты основного тона от одного питча к другому – $T_{0\max}$.

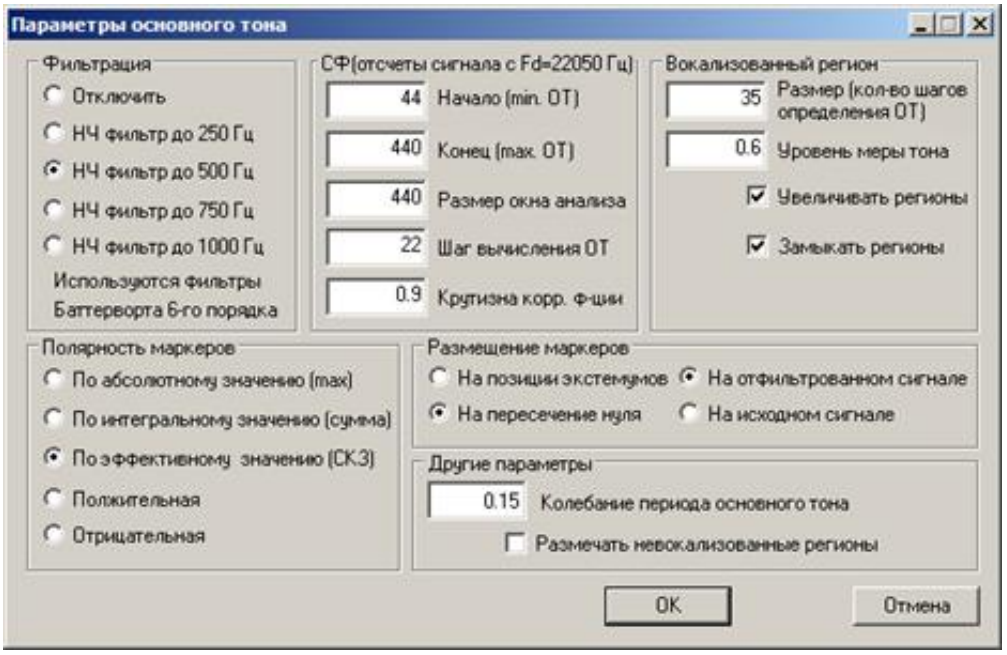


Рис. 5. Диалог настройки модуля разметки

Тестирование точности расстановки маркеров основного тона речевого сигнала проводилось на записях голосов трех дикторов: двух мужских и одном женском. Каждым диктором были надиктованы 500 тестовых фраз [7]. Каждая из надиктованных фраз была разсегментирована на аллофоны. Определение точности расстановки маркеров ОТ выполнялось для следующих групп аллофонов:

1 ударные гласные A0,O0,E0,I0,Y0;
2 гласные первой степени редукции A1,O1,E1,I1,Y1;
3 гласные второй степени редукции A2,O2,E2,I2,Y2;
4 сонорные носовые M,N, M',N';
5 сонорные L, L',V,R,J',V',R';
6 звонкие щелевые Z,Z',Zh;
7 звонкие взрывные P,D,G,P',D',G'.

Определение точности установки питчей для каждой группы аллофонов выполнялось путем вычисления для каждого аллофона средней частоты основного тона и определения количества питчей, попавших в диапазон допустимого отклонения текущего значения ЧОТ от среднего. Результаты экспериментальных исследований для различных диапазонов допустимого отклонения приведены в табл. 1.

	Допустимое отклонение	Группа 1	Группа 2	Группа 3	Группа 4	Группа 5	Группа 6	Группа 7
Диктор 1 (М)	2 раза	100.00%	99.99%	100.00%	100.00%	99.98%	100.00%	100.00%
	1.5 раза	99.94%	99.95%	100.00%	99.94%	99.58%	100.00%	99.85%
	1.3 раза	99.53%	99.20%	99.81%	99.55%	97.19%	99.15%	98.16%
Диктор 2 (М)	2 раза	99.99%	99.99%	100.00%	100.00%	99.93%	100.00%	100.00%
	1.5 раза	99.94%	99.96%	99.90%	99.90%	99.14%	99.83%	99.49%
	1.3 раза	99.66%	99.48%	99.72%	99.36%	95.44%	99.07%	97.84%
Диктор 3 (Ж)	2 раза	99.99%	99.98%	100.00%	99.97%	99.94%	99.99%	99.95%
	1.5 раза	99.88%	99.63%	99.97%	99.54%	98.91%	99.77%	99.14%
	1.3 раза	99.07%	98.83%	99.58%	98.69%	95.05%	98.91%	96.76%

Таблица 1. Результаты исследования точности установки питчей

Заключение

Приведенная система разметки речевого сигнала на периоды основного тона использована для подготовки баз данных речевых элементов для систем компиляционного синтеза речи по тексту. Она может быть так же использована в системах распознавания речи и верификации диктора.

Список литературы

1. Lobanov B.M., Tsyrulnik L.I. Phonetic–Acoustic Problems of Personal Voice Cloning by TTS //Proceedings of the Ninth International Conference “Speech and Computer” SPECOM’2004, Saint-Petersburg, 2004 – pp. 17–21.

2. Charpentier F., Moulines E. Pitch Synchronous Waveform Processing Techniques for TTS Synthesis using Diphones //Proceedings of Eurospeech’89, Paris, 1989 – pp. 13–19.

3. Hess W. Pitch determination on Speech Signals with Special Emphases on Time-Domain Methods. Proc. of NCVS Workshop on Voice Analysis, The Center of Performing Arts, Denver, February 1994.

4. Бабкин В.В. Помехоустойчивый выделитель основного тона речи //7-я Международная Конференция и Выставка Цифровая Обработка Сигналов и ее Применение DSPA-2005 – Москва 16-18 марта 2005 г.

5. Valery A. Petrushin Adaptive Algorithms for Pitch-synchronous Speech Signal Segmentation //Proceedings of the Ninth International Conference “Speech and Computer” SPECOM’2004, Saint-Petersburg, 2004 – pp. 146–153.

6. Л.Р. Рабинер, Р.В. Шафер Цифровая обработка речевых сигналов: Пер. с англ./Под ред. М.В.Назарова и Ю.Н.Прохорова – М.: Радио и связь, 1981 – 496 с., ил.

7. Передача речи по трактам радиотелефонной связи. Требования к разборчивости речи и методы артикуляционных измерений: ГОСТ 16600-72. –Введ. 27.09.1972. – Москва: Государственный комитет стандартов Совета Министров СССР, 1973. – 90 с.

---

[1] *Работа выполнена при поддержке европейского фонда INTAS в рамках проекта «Разработка многоголосовой и многоязыковой системы синтеза и распознавания речи (языки: белорусский, польский, русский)» в соответствии с грантом INTAS № 04-77-7404.*